International Journal of Trend in Scientific Research and Development (IJTSRD)
Conference Issue | March 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Fostering Innovation, Integration and Inclusion Through Interdisciplinary Practices in Management

Machine Learning Approach for Employee Attrition Analysis

Dr. R. S. Kamath[1], Dr. S. S. Jamsandekar[2], Dr. P. G. Naik[3]

[1]Associate Professor, [2]Assistant Professor, [3]Professor
[1,2,3]Department of Computer Studies,
[1,2,3]Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, Maharashtra, India

Organised By:
Management Department, Chhatrapati Shahu Institute of Business Education and Research, Kolhapur, Maharashtra

An Autonomous Institute Under UGC & Shivaji University
College with Potential for Excellence (CPE) - III Phase.

How to cite this paper: *Dr. R. S. Kamath | Dr. S. S. Jamsandekar | Dr. P.G. Naik "Machine Learning Approach for Employee Attrition Analysis" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Special Issue | Fostering Innovation, Integration and Inclusion Through Interdisciplinary Practices in Management, March 2019, pp.62-67, URL: https://www.ijtsrd.com/papers/ijtsrd23065.pdf*

IJTSRD23065

ABSTRACT

Talent management involves a lot of managerial decisions to allocate right people with the right skills employed at appropriate location and time. Authors report machine learning solution for Human Resource (HR) attrition analysis and forecast. The data for this investigation is retrieved from Kaggle, a Data Science and Machine Learning platform [1]. Present study exhibits performance estimation of various classification algorithms and compares the classification accuracy. The performance of the model is evaluated in terms of Error Matrix and Pseudo R Square estimate of error rate. Performance accuracy revealed that Random Forest model can be effectively used for classification. This analysis concludes that employee attrition depends more on employees' satisfaction level as compared to other attributes.

INTRODUCTION

The process to identifying the existing talent in an organization is among the top talent management challenges and the important issue. For every organization, human resource plays a vital role in all strategic decisions. Satisfied, highly-motivated and loyal employees represent the basis of a company and which in turn have impacts on the productivity of an organization.

The prime objective of the present study is to analyze why some of the best and most experienced employees are leaving prematurely. This analysis also wishes to predict which valuable employees will leave next.

The rest of paper is designed as follows; Introduction followed by the materials and methods utilized in the present study. Then the third section summarizes the results and discussions of the HR attrition analysis. The conclusion at the end justifies the suitability of Random Forest model for this talent mining.

Materials and Methods

The dataset for the present analysis is taken from Kaggle, Machine Learning platform [1]. This is the simulated dataset comprising 15000 employee records classified into two categories (left or not left) based on satisfaction level, latest evaluation, number of project worked on, average monthly hours, time spend in the company, work accident, promotion within the past 5 years, department and salary. Table 1 gives description of employee dataset.

Table 1: Employee dataset description for talent mining

Attribute	Description	Data Type
satisfaction_level	Level of satisfaction (0-1)	Numeric
last_evaluation	Time since last performance evaluation (in Years)	Numeric
number_project	Number of projects completed while at work	Numeric
average_montly_hours	Average monthly hours at workplace	Numeric
time_spend_company	Number of years spent in the company	Numeric
Work_accident	Whether the employee had a workplace accident	Numeric
Left	Whether the employee left the workplace or not (1 or 0)	Numeric
promotion_last_5years	Whether the employee was promoted in the last five years	Numeric
sales	Department in which they work for	String
salary	Relative level of salary (high)	String

This section explores details of experiment conducted for employee attrition analysis and forecasting. The present study is carried out using R and Rattle data mining platform [4]. Figure 1 shows summary of the HR dataset. Dataset is partitioned randomly into training, testing and validation with division 70%, 15 % and 15% respectively. We used the training dataset for parameter adjustment of model whereas validation set to control learning process.

```
satisfaction_level last_evaluation  number_project  average_montly_hours
Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
Median :0.6500     Median :0.7200   Median :4.000   Median :200.0
Mean   :0.6159     Mean   :0.7152   Mean   :3.798   Mean   :200.9
3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0

time_spend_company Work_accident    promotion_last_5years         sales
Min.   : 2.000     Min.   :0.0000   Min.   :0.00000       sales      :2897
1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.00000       technical  :1915
Median : 3.000     Median :0.0000   Median :0.00000       support    :1580
Mean   : 3.498     Mean   :0.1446   Mean   :0.02257       IT         : 852
3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.00000       product_mng: 612
Max.   :10.000     Max.   :1.0000   Max.   :1.00000       marketing  : 606
                                                          (Other)    :2037

   salary          left
high  : 842   Min.   :0.0000
low   :5090   1st Qu.:0.0000
medium:4567   Median :0.0000
              Mean   :0.2367
              3rd Qu.:0.0000
              Max.   :1.0000
```

Figure 1: Dataset exploration – Summary

Among the vast machine learning algorithms, authors have picked Decision Tree, Random Forest, Support Vector Machine (SVM), and Linear Regression techniques to build the model. These algorithms are based on supervised learning and best known for building prediction models [8]. Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features/ predictors such that we can predict the output values for new data based on those relationships which it learned from the previous data sets.

Figure 2 explains Decision tree modeling of HR data. It begins with a root node "satisfaction level", that part into different branches, prompting to further nodes, each of which may additionally part or else end as a leaf node. Connected with each nonleaf node will be a test or question that figures out which branch to take after [7]. The leaf nodes indicate the attrition sates whether the employee "left" or "not left". Figure 3 gives pictorial representation of Decision tree thus derived.

```
Classification tree:
rpart(formula = left ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] average_montly_hours last_evaluation      number_project
[4] satisfaction_level   time_spend_company

Root node error: 2485/10499 = 0.23669

n= 10499

        CP nsplit rel error  xerror      xstd
1 0.240644      0   1.00000 1.00000 0.0175262
2 0.184909      1   0.75936 0.75936 0.0158321
3 0.073843      3   0.38954 0.38954 0.0119291
4 0.054728      5   0.24185 0.24185 0.0095788
5 0.031388      6   0.18712 0.18833 0.0085093
6 0.016901      7   0.15573 0.15976 0.0078650
7 0.010060      8   0.13883 0.14165 0.0074223
8 0.010000      9   0.12877 0.13119 0.0071521

Time taken: 0.39 secs
```

Figure 2: Decision tree modeling

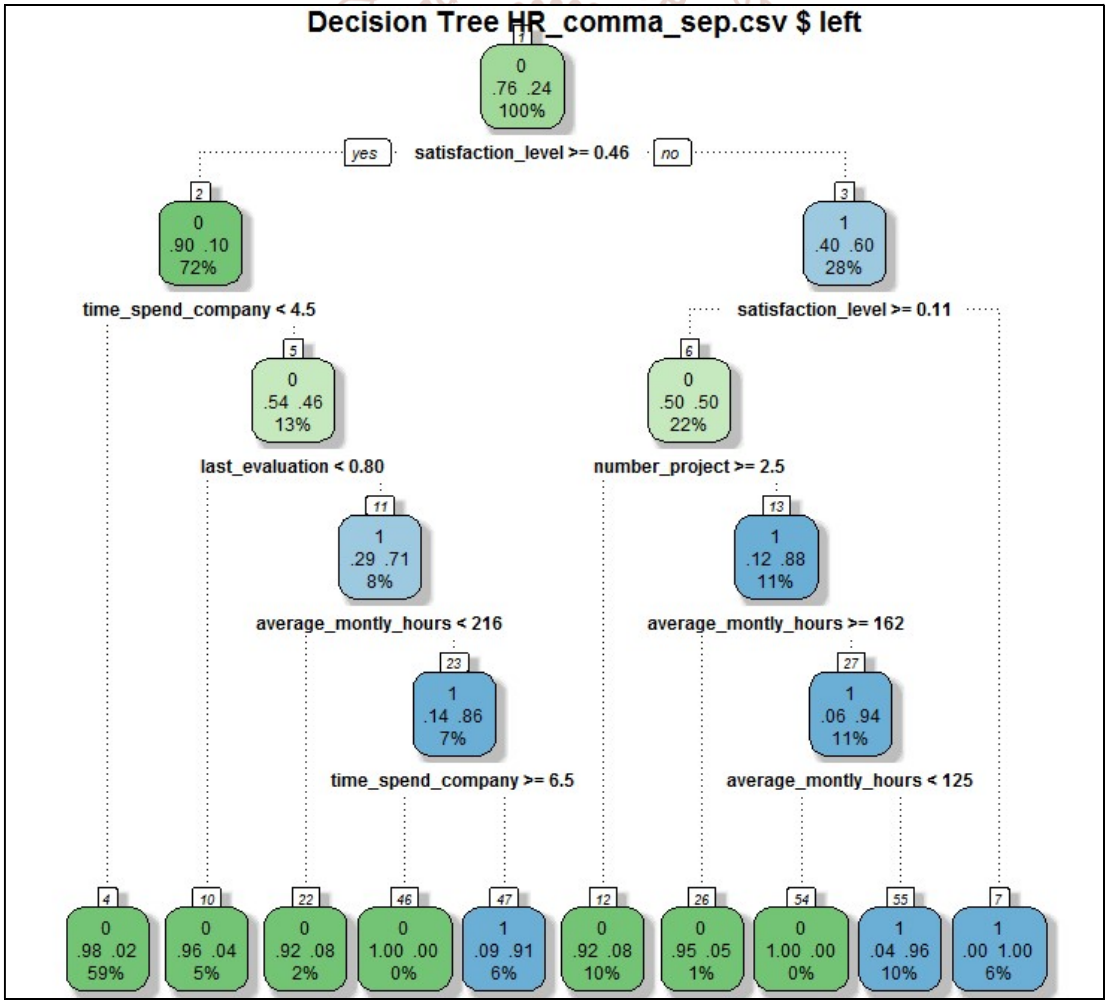

Figure 3: Decision tree for HR attrition status

Figure 4 explains Random Forest Modeling for HR attrition analysis. RANDOMFOREST package in R environment is employed here to analyze model structure [5-6]. RF builds many decision trees using random subset of data and variables. Rattle provides access to three parameters such as the number of trees, sample size and number of variables for tuning the models.

```
Number of observations used to build the model: 10499
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(left) ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 200, mtry = 3, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.9%
Confusion matrix:
     0    1 class.error
0 8002   12  0.00149738
1   82 2403  0.03299799
```

Figure 4: Summary of the Random Forest Model

Figure 5 explains Support Vector Machine (SVM) designed for the attrition analysis of employee data. SVM searches for support a vector that separates the class.

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.109166570184432

Number of Support Vectors : 1510

Objective Function Value : -1188.127
Training error : 0.03467
Probability model included.

Time taken: 10.80 secs
```

Figure 5: Summary of SVM Model

Figure 6 explains Linear Regression Model. It is the traditional method for fitting a statistical model to data. It is appropriate since the target variable "attrition status" is numeric.

```
Call:
glm(formula = left ~ ., family = binomial(link = "logit"), data = crs$dataset[crs$train,
    c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2650  -0.6586  -0.4000  -0.1137   3.1235
```

Figure 6: Summary of Logistic Regression model

Results and Discussion

The present investigation employed different prediction algorithms to analyze employee attrition status and likelihood of retention-attrition of employees. The performance of the model is evaluated in terms of Error Matrix and Pseudo R Square estimate of error rate. An error matrix shows the true outcomes against the predicted outcomes. It is also known as confusion matrix. Table 2 explains performance analysis of these classifiers in terms of error matrix.

Table 2: Performance Analysis of the Classifiers

Model	Error Matrix
Decision Tree	Predicted Actual 0 1 Error 0 1662 21 1.2 1 41 525 7.2
Random Forest	Predicted Actual 0 1 Error 0 1680 3 0.2 1 15 551 2.7
Support Vector Machine	Predicted Actual 0 1 Error 0 1641 42 2.5 1 50 516 8.8
Liner Model	Predicted Actual 0 1 Error 0 1558 125 7.4 1 396 170 70.0

Figure 7, the "Predicted versus Observed" plot shows the performance analysis of all the four models. The plot displays the predicted values against the observed values. The Pseudo R-Squared, square of the correlation between the predicted and observed values. The closer to 1, is the acceptable one. Table 3 gives Pseudo R-Square values for these four models.

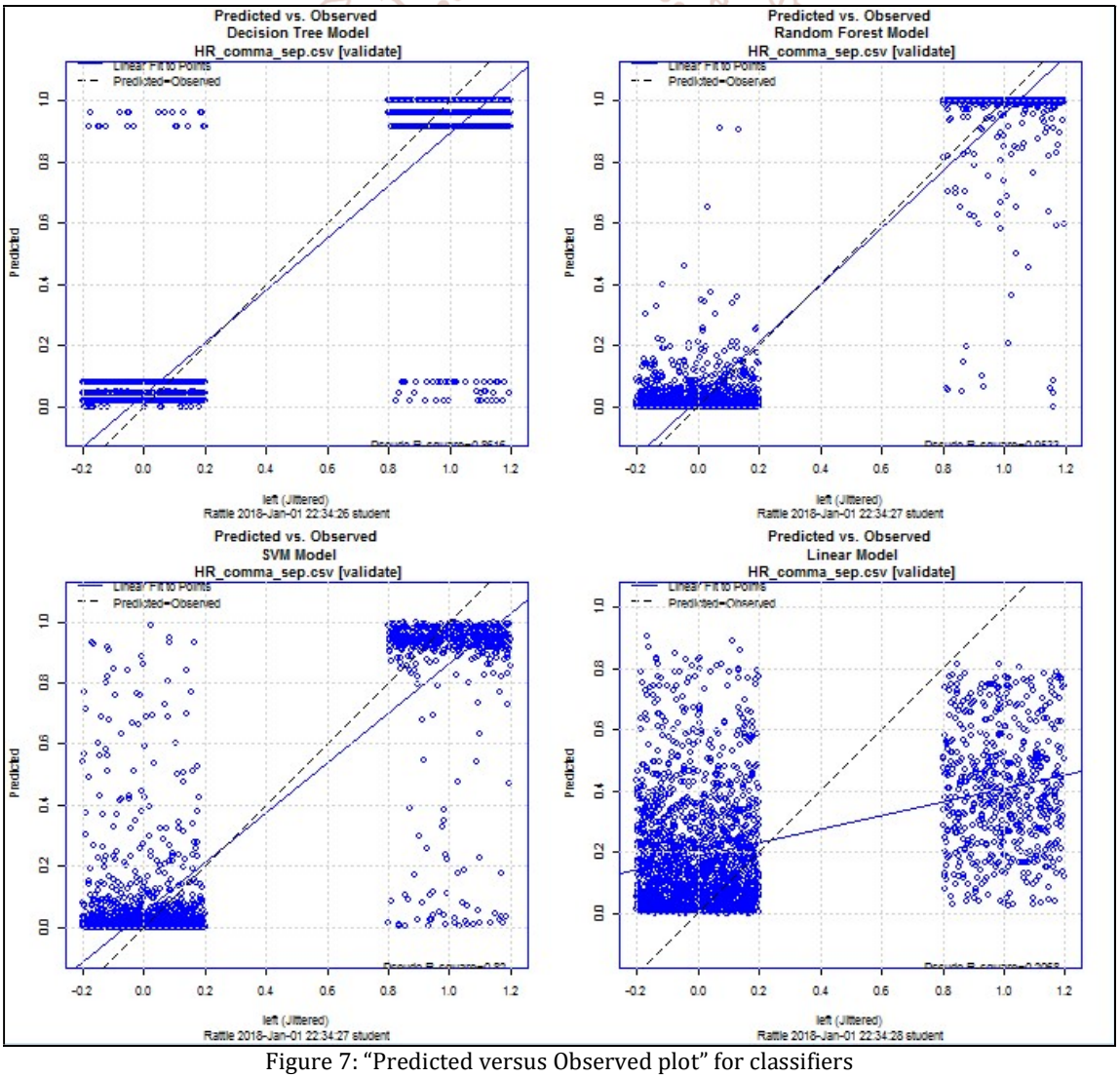

Figure 7: "Predicted versus Observed plot" for classifiers

Table 1: Performance accuracy of classifiers

Classifier	Pseudo R-square
Decision Tree	0.8473
Random Forest	0.9773
Support Vector Machine	0.8315
Linear Regression	0.2299

Confusion matrix and "Predicted versus Observed" plot concludes that Random Forest is the appropriate model for analysis of Employee attrition as compared to the other algorithms considered in this study and the underlined data. Figure 8 explains the relative importance of HR dataset attributes using Gini importance and Permutation importance measures. Based on these two measures, it reveals that employees' "satisfaction level" is the predominant predictor of employee attrition.

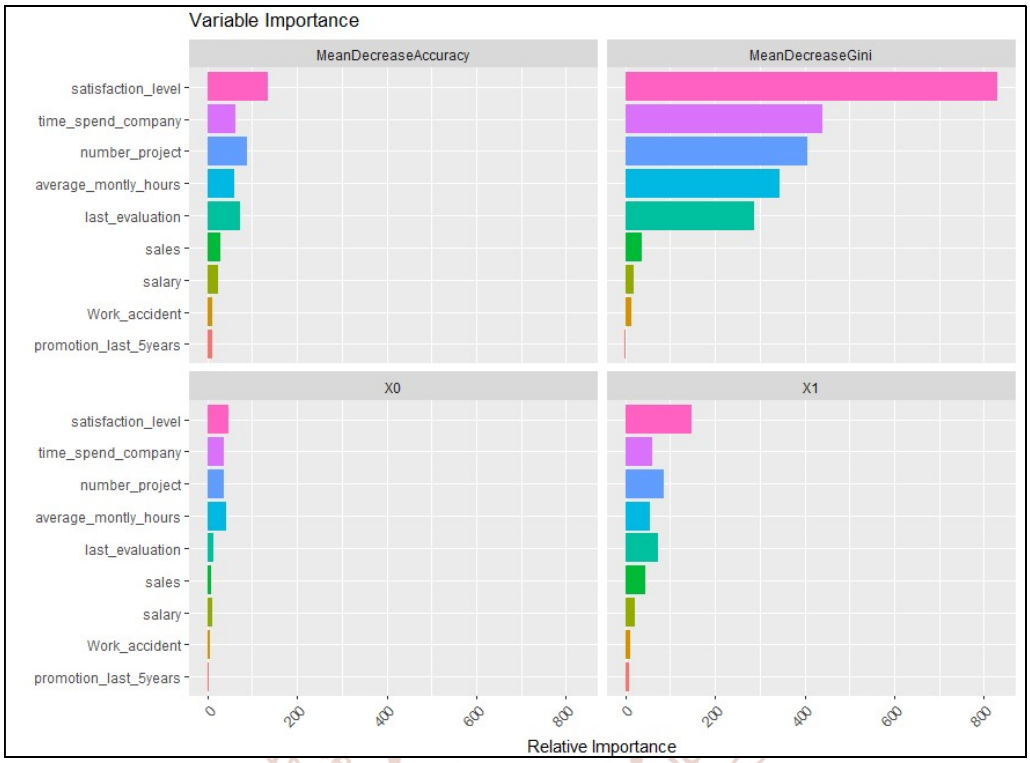

Figure 8: Dependency of employee attrition status on other attributes

Conclusion

Authors have explored a machine learning solution for HR attrition analysis and forecast. Present study exhibits performance estimation of various classification algorithms and compares the classification accuracy. The performance of the model is evaluated in terms of Error Matrix and Pseudo R Square estimate of error rate. Performance accuracy revealed that Random Forest model can be effectively used for classification. The result also concludes that employee attrition depends more on employees' satisfaction level as compared to other attributes.

References:

[1] Retrieved on 30th Dec, 2017 from https://www.kaggle.com/ludobenistant/hr-analytics-1/data

[2] Boudreau, J. W. – Ramstad, P. M.: Beyond HR. Boston. Harvard Business School Press, 2007. ISBN 978-1-4221-0415-6.

[3] https://www.infogix.com/blog/machine-learning-vs-statistical-modeling-the-real-difference, accessed date 28/012/2017

[4] Graham, W. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Springer, DOI 10.1007/978-1-4419-9890-3

[5] Breiman, L. (2005), Random Forest. Machine Learning,45, 5-32

[6] Andy, L., & Matthew, W. (2002). Classification and Regression by random Forest, R News, 2(3)

[7] R. S. Kamath, R. K. Kamat (2016), Modeling of Random Textured Tandem Silicon Solar Cells Characteristics: Decision Tree Approach, Journal of Nano and Electronic Physics, Vol. 8 No 4(1), 04021(4pp)

[8] R. S. Kamath, R .K. Kamat (2016), Supervised Learning Model for Kick starter Campaigns with R Mining, International Journal of Information Technology, Modeling and Computing (IJITMC), Vol. 4, No.1, February, 19-30